Table 1
An Example of Interconnected Traffic Flows

Assumptions For the Example

100 Total Subscribers
90 subscribers to the LEC
10 subscribers to the CMRS provider

Each subscriber calls every other subscriber once each month

Calls per Month

Each CMRS Subscriber			
Calls to LEC subscribers	90	Calls from LEC subscribers	90
Calls to CMRS subscribers	9	Calls from CMRS subscribers	9
Total calls made	99	Total calls received	99
Each LEC Subscriber			
Calls to LEC subscribers	89	Calls from LEC subscribers	89
Calls to CMRS subscribers	10	Calls from CMRS subscribers	10
Total calls made	99	Total calls received	99

Total Calls Terminated

By the CMRS Provider:
90 calls per CMRS subscriber x 10 subscribers = 900 calls

By the LEC:
10 calls per LEC subscriber x 90 subscribers = 900 calls

A consumer's decision to subscribe to a service depends on the value expected from both placing and receiving calls.[6] The example illustrates how, for CMRS providers, a high proportion of calling and thus of the value of their service to consumers depends on interconnection with the LEC network. A much smaller proportion of the calling for LEC subscribers, and thus much less of the value of LEC service to a subscriber, depends on interconnection to the CMRS provider. As a result of this

[6] More formally, the demand for access (subscription) depends on the consumer surplus received from both originated and received calls, after taking into account charges for usage. See L. Taylor, *Telecommunications Demand in Theory and Practice*, Kluwer, Dordrecht, 1994.

asymmetry, the LEC can expect to be in a far stronger bargaining position. The LEC "needs" interconnection less than the CMRS provider, and can far more credibly threaten to "walk away" from the bargaining table if it doesn't get what it wants. By charging CMRS providers a high price for interconnection, the LEC can use its market power over this input to extract supracompetitive profits from downstream markets in which CMRS services are sold. The imbalance of bargaining power indicates that LECs likely will be able to exercise such market power.

When CMRS providers and LECs compete downstream in the supply of retail telecommunications services, a second economic factor will affect the outcome of negotiated interconnection agreements. LECs likely will have an incentive to disadvantage the competing CMRS provider in order to increase or preserve the market power they can exercise in downstream markets.[7,8] Forcing the CMRS provider to pay still more for interconnection reduces the competitive pressure the CMRS provider can exert on the LEC in downstream markets; in the extreme high interconnection charges could drive CMRS competitors out of business.[9] The result is less downstream competition from which consumers can benefit. The potential for disadvantage is clear if

---

[7] As is well known, under the right conditions an upstream supplier with monopoly control of an input can capture monopoly profits as effectively as if it had a downstream monopoly. In such cases it has no incentive to acquire a monopoly downstream. The conditions necessary for this result include that the input is used in fixed proportion to output, that the downstream market is perfectly competitive, and that there are no regulatory constraints on the input price. See for example, Michael H. Riordan and Steven C. Salop, "Evaluating Vertical Mergers: A Post-Chicago Approach," *Antitrust Law Journal*, Vol. 63, No. 2 (Winter 1995). It is unlikely that all the necessary conditions are satisfied in this case.

[8] A LEC need not be pricing local services above its own costs, and in particular need not be pricing to recover revenue in excess of accounting costs, in order to have an incentive to disadvantage rival CMRS providers. LECs could have such a incentive so long as competition from the rival would, with lower interconnection prices, reduce or constrain LECs' pricing and/or the net revenue they can earn in the downstream market. Such pressure on LEC prices and/or net revenues might result from rivals simply being able to offer lower prices that LECs must match to avoid losing market share, or from rivals offering service of superior quality that LECs must match or offset with lower prices to avoid losing market share.

[9] LECs may also be able to disadvantage rival CMRS providers through non-price as well as price terms.

the LEC can charge CMRS providers more per call or per minute of use to terminate calls on the LEC network than the LEC pays for termination by the CMRS provider.[10]

Even if a LEC must pay the same price for termination that it charges a CMRS provider, it still may be able to disadvantage a competing CMRS rival. When CMRS providers are just beginning to compete with LECs, termination on the other network is likely to be a more important input per subscriber and to account for a larger proportion of costs for CMRS providers than for LECs (as in the example above). Even if both pay the same price for termination, an increase in that price raises costs per subscriber more for the CMRS provider than for the LEC, and thus disadvantages the CMRS provider relative to the LEC.

Because of the unequal bargaining positions of the parties, and because of the incentive of LECs to use pricing of interconnection service to extend their market power, this analysis suggests that private, unconstrained negotiations between LECs and CMRS providers are unlikely to yield efficient interconnection compensation arrangements that are in consumers' interests. Many existing interconnection arrangements between cellular providers and LECs have been the result of negotiations subject to little or no regulatory oversight. The CTIA has collected information from members on what they pay LECs for terminating traffic, and on what (if anything) they are paid to terminate LEC-originated calling. The information collected is consistent with the analysis here. All of the cellular systems responding to this question reported that they must pay LECs to terminate traffic originated by cellular subscribers. Few cellular systems, however, receive compensation for terminating calls placed by LEC subscribers; only 10 percent of members' responses indicate that they receive any compensation from LECs for terminating LEC-originated traffic -- despite the FCC policy requiring mutual compensation.[11] Several cellular systems reported that they not only failed to receive compensation, they in fact had to *pay* the LEC for LEC-originated traffic.

---

[10] If termination costs differ for the two networks, there would be a disadvantage if the markup of price over cost is greater for calls terminated by the LEC.

[11] The reported figure of 10 percent is in fact something of an overstatement. Each of the responses indicating compensation was received is the response of a cellular operator providing

If compensation arrangements for interconnection are not to be left to unconstrained private negotiations, policy choices must be made among alternative ways of structuring and setting the level for compensation. The remainder of this paper discusses issues raised by such choices. The next section discusses whether and how various compensation arrangements allow carriers to recover the costs the incur as a result of interconnection. The remaining sections discuss the impact on economic efficiency of the structure and level of rates under various compensations arrangements.

III. Recovery of Costs

Two interconnected carriers will each incur costs to handle traffic originated on one network and terminated on the other. Most of these costs will be incurred to terminate traffic originated on the other's network. Other interconnection-related costs include those of the trunks connecting the two networks, and monitoring, billing and accounting costs. The first characteristic of compensation arrangements to be evaluated is whether they allow for the recovery of such costs. Does each carrier, as a result of the interconnection agreement, incur an obligation to pay the additional costs incurred to handle interconnected traffic? This section discusses the cost recovery characteristics of usage sensitive payments and bill and keep compensation arrangements.

A. Recovery of Costs with Usage Sensitive Payments

LECs and CMRS providers have a direct means for recovering the cost of terminating traffic originated by the other if the compensation arrangement specifies that each will make payments tied to the volume of terminated traffic. Whether those payments are adequate to cover the costs incurred, both now and in the future, will depend on the level and structure of the usage sensitive rates. Rate level and rate structure issues are discussed in more detail in subsequent sections, but the basic implications for cost recovery are straightforward.

---

information for multiple systems, and each response indicates that compensation is received for only *some* of the systems covered by the response.

If the volumes of interconnected traffic were constant, it would be a simple matter, in principle, to determine rates per unit of traffic that would recover those costs. First, determine the level of costs per billing period that each carrier incurs to terminate traffic originating on the network of the other, and then divide by the number of units of traffic terminated in the billing period.[12] Charging this rate will yield revenues equal to the costs used to calculate the rate, so long as traffic volumes do not change. Of course, in practice traffic volumes will change over time, both because the rate the other carrier pays for interconnection and includes in retail prices will affect consumers' usage, and because traffic can be expected to change over time with changes in the number of subscribers to each network and with the overall growth in demand for various types of local service.

Whether, in the face of changing traffic, a given set of rates will continue to generate revenue equal to, or greater than, the costs incurred, will depend on how closely the structure and level of rates match the structure and level of costs. When traffic increases, revenue will grow more rapidly than cost if there is a uniform charge per minute of terminated traffic and the cost of terminating traffic does not increase at the same rate as traffic. Conversely, with this structure of rates and costs a reduction in traffic would cause revenue to fall more rapidly than cost.

The structure of usage sensitive rates almost certainly will diverge from the structure of costs if such rates are used to recover costs that do not depend directly on the level of usage. Even for rates used to recover only usage-sensitive costs, revenue and cost may diverge over time. Costs may vary with usage without varying in proportion to total traffic. Later sections of this paper discuss the impact of rate structure and level on overall economic efficiency. The point here is that differences between the structure of rates and the structure of costs can, over time, lead to growing differences between the revenue a carrier recovers and the costs it incurs.

[12] The derivation of rates from costs is discussed in more detail below.

B. Recovery of Costs Under Bill and Keep

If a LEC and a CMRS provider interconnect under a bill and keep arrangement, neither makes any payments to the other. Instead, each carrier must cover the cost of handling interconnected traffic by billing its own customers and keeping the revenue. It is often observed that when traffic flows between the carriers are balanced the net flow of revenue between carriers is the same under both a bill and keep system and under uniform usage payments. This is a useful and significant observation. A more complete comparison of bill and keep and usage sensitive payments, however, requires an analysis of the costs the two providers incur to handle interconnected traffic, and the obligations each incurs to pay for those costs.

1. LEC and CMRS Cost Obligations

Calling bill and keep a compensation arrangement may seem a misnomer as neither the LEC nor the CMRS provider makes payments to compensate the other for costs incurred. This does not mean, however, that either carrier receives interconnection services for free. The LEC and the CMRS provider each incur a cost obligation in exchange for the interconnection services they receive from the other. Bill and keep is part of a mutual obligation to terminate traffic from the other. The CMRS provider receives termination services from the LEC only in exchange for providing, and bearing the costs of providing, termination services for LEC-originated traffic. The LEC likewise receives termination services from the CMRS provider in exchange for providing similar services for CMRS-originated traffic.

Under bill and keep each provider incurs a cost obligation in exchange for receiving interconnection services from the other. However, its costs may, or may not, be equal to what it costs the other provider to provide interconnection services. Under bill and keep, when each provider must incur approximately the same costs to supply interconnection services to the other, the cost a provider incurs will equal the cost of the service it receives.

2. ***Balance of Traffic***

Is balance in the traffic flows between a LEC and CMRS provider equivalent to a balance in the costs of supplying interconnection services? This is not necessarily the case. The traffic flows in each direction need not be equal for each provider to bear costs approximately equal to the cost of interconnection services they receive, nor are equal traffic flows sufficient to insure that the costs are equal. Equality of overall traffic flows between the providers is neither necessary nor sufficient for equality of costs because the impact of traffic on cost varies with the time of day, and because LEC and CMRS networks may incur different costs per unit of terminated traffic.

The "usage-sensitive" costs of terminating traffic in a LEC or CMRS network arise from the need to provide sufficient network capacity to carry any additional traffic that is terminated during the busy hour in that network. Each network's switching and transport facilities are sized to provide a specified grade of service at the busy hour. Once each network has been constructed, nearly all of the incremental costs of carrying additional traffic are due to expanding capacity. Because the principal costs of terminating traffic are capacity costs, the hourly distribution through the day of terminating traffic is central to determining the effect of terminating traffic on network costs. The volume of traffic delivered during the *terminating* network's busy hour will determine the costs of providing terminating interconnection services, since this traffic will affect the capacity needed by the terminating network.[13] Terminating traffic delivered outside the busy hour will have little effect on needed capacity and therefore little effect on costs. Thus, it is the balance in the amount of traffic delivered to each provider during its busy hour that will affect costs rather than the balance of overall traffic.

The balance of overall traffic could, for a variety of reasons, differ substantially from the balance of traffic delivered during the terminating carrier's busy hour. The hypothetical examples summarized in Table 2 illustrate two reasons the patterns could

---

[13] This is something of a simplification since not all network facilities necessarily have the same busy hour.

differ. In Case A, both the LEC and CMRS provider have the same busy hour, and both deliver 100 units of traffic to the other during that busy hour. The time profile of LEC-originated and CMRS-originated calling differ, however. Over a 24-hour period total LEC-originated, interconnected traffic is 8 times the amount of traffic in the busiest hour for such traffic, while total CMRS-originated, interconnected traffic is 12 times the amount in the busiest hour. As a result, the LEC terminates 1200 units of total traffic during a 24-hour period, while the CMRS provider terminates a total of 800 units. Despite this imbalance in total traffic, each provider terminates the same amount of traffic during the terminating system's busy hour, suggesting that each much provide

Table 2
Balance of Traffic:
Total Traffic and Traffic Imposing Capacity Costs

	Direction of Traffic	
	CMRS to LEC	LEC to CMRS
Hypothetical Case A		
Terminating system busy hour (BH)	11am	11am
Terminating traffic in terminating system BH	100	100
Ratio of total interconnected traffic to BH traffic	12	8
Total 24- hour interconnected traffic	1200	800
Hypothetical Case B		
Terminating system busy hour (BH)	11am	4pm
Terminating traffic in terminating system BH	100	100
Maximum volume of traffic in any hour	125	100
Ratio of total interconnected traffic to max. hourly traffic	10	10
Total 24- hour interconnected traffic	1250	1000

approximately the same increment in overall capacity to handle interconnected traffic.

The hypothetical of Case B illustrates another possible source of difference: non-coincident busy hours for the two networks with interconnected traffic. In Case B, the LEC and CMRS provider again each receive 100 units of traffic for termination during their system busy hour. Now, however, the two networks have different system busy hours, and the system busy hour for the LEC is not the hour during which it receives the most terminating traffic. During some other hour (perhaps the CMRS busy hour), the LEC receives 125 units of traffic. If the ratio of total traffic to busiest hour traffic is 10 for traffic originated on the each network, total CMRS-originated traffic would be 1250 units, and total LEC-originated traffic 1000 units. Again in Case B, the LEC terminates more total traffic during a 24-hour period, but would not have to add more capacity for terminating traffic than would the CMRS provider.[14]

3. LEC-CMRS Traffic Patterns

To obtain factual information on the time profile of CMRS traffic, and on interconnected traffic between LECs and CMRS providers, the CTIA has collected data from member systems.

The collected information shows, as expected, that the amount of traffic carried on cellular systems varies throughout the business day and has a pronounced peak. A composite traffic profile for surveyed systems reporting hourly traffic patterns is shown in Figure 1.[15]

This composite traffic profile shows an overall busy hour peak for traffic from 4-5 PM. This composite result is consistent with the busy hours reported for cellular systems. A total of 51 percent of responses reported a cellular system busy hour of 4-5 PM (and an additional 20 percent reported a system busy hour of 5-6 PM). Survey information, although fragmentary on this point, suggests that the cellular system busy hour of 4-5 PM also may be the hour during which LECs deliver the most traffic for termination.[16] If this is accurate, cellular systems must terminate the largest volume of LEC-originated traffic during the cellular system busy hour.

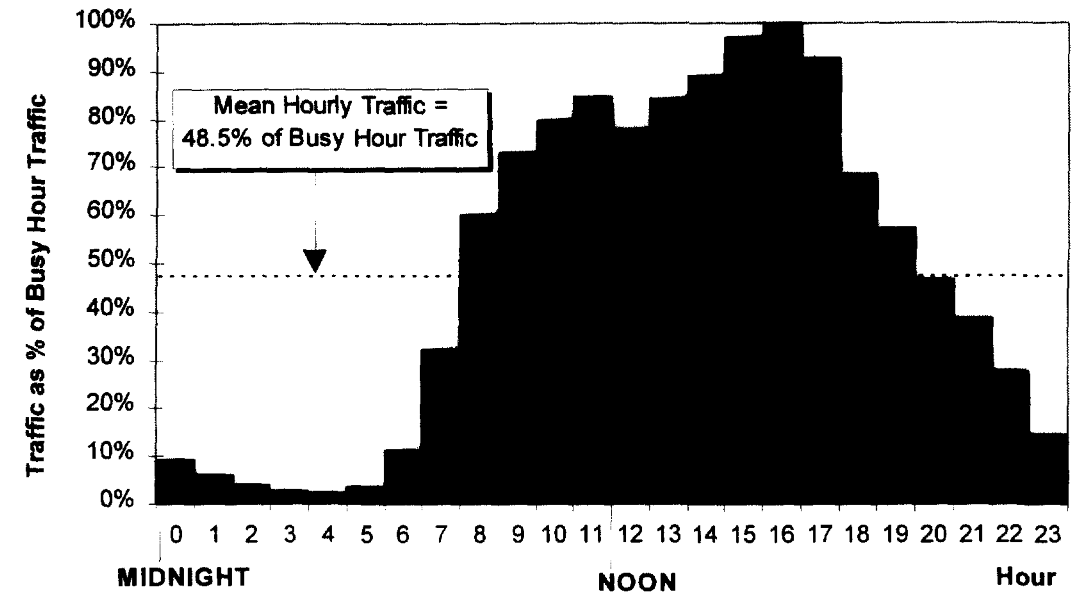
---

[14] This assumes that the amount of terminating traffic received by the LEC is not sufficiently great to shift its system busy hour.

[15] The data from which this traffic profile was calculated were for average business day traffic.

[16] One explanation for this pattern would be that the time profile both of calls both placed and received by cellular subscribers is strongly influenced by when subscribers are in their cars or have their portable units tuned on.

Figure 1
Composite Traffic Profile of Responding Cellular Systems



Source: Survey of CTIA members and CRA calculations.

This survey information suggests that the busy hours of cellular systems and LECs often will not be the same. Our understanding is that the busy hour for many LEC facilities, and often the system busy hour, is in the late morning or early afternoon, rather than the later afternoon. Only 2 percent of responses reported a cellular systems busy hour between 10 AM and noon, and only 5 percent reported a cellular system busy hour between 1 PM and 3 PM. If this is accurate, the traffic delivered to a LEC for termination would be at a maximum for many LECs outside their system busy hour, assuming, as seems likely, that the cellular system busy hour of 4-5 PM is also the hour when cellular systems deliver the most traffic to LECs for termination. LECs would receive a smaller volume of traffic for termination during their busy hour. The traffic profile in Figure 1 shows the volumes of cellular system traffic at 11 AM - noon and 2-3 PM are roughly 85 and 89 percent as large as traffic volumes during the cellular system busy hour. As in the hypothetical example in Case B above, this pattern also would make the amount of traffic each must terminate during its system busy hour more nearly balanced than the flow of total traffic.

In the composite traffic profile of Figure 1, mean hourly traffic over the business day is slightly less than half as great as busy hour traffic, and total traffic during the business day is about 11.6 times busy hour traffic. Traffic profiles for LECs show a ratio of busy hour traffic to mean hourly traffic of about 2.5 - 3, implying that total traffic is roughly 8 - 10 times as large as busy hour traffic.[17] If the ratio of total calling to busiest hour calling is the same for LEC-originated traffic delivered to cellular systems as for all LEC calling, the pattern would be similar to that hypothesized in Case A above. This pattern would imply that the imbalance between total cellular-originated calling and LEC-originated calling would be greater than the imbalance in traffic terminated during each system's busy hour. Unfortunately, it was not possible to collect information on the time profile of LEC to mobile calling that could provide direct confirmation (or refutation) of the existence of this pattern.

In present day cellular systems, the time during which LEC subscribers can reach a CMRS handset often is limited, either by the amount of time cellular subscribers are near their cars or by the battery life of portable handsets. These factors, among others, result in an imbalance between total LEC to cellular and cellular to LEC traffic. It was possible to collect information from CTIA members only on the relative amount of traffic to and from LECs over a 24 hour period, but not on traffic received for termination during the busy hour of each network. Based on responses that provided sufficient data for the calculation, cellular systems on average received from LECs and terminated about a third as much total traffic as LECs received from CMRS providers and terminated.[18]

---

[17] Rolla E. Park, *Incremental Costs and Efficient Prices with Lumpy Capacity: The Two Product Case*, The Rand Corporation, Santa Monica, 1994.

[18] The reported figure is calculated from the means of responses to questions requesting the percent of cellular system traffic with various originating and terminating patterns. LEC-terminated traffic here does not include traffic passed on to IXCs (or traffic direct trunked to IXCs by cellular systems); LECs receive switched access payments from IXCs for such traffic , and this traffic may be less costly to terminate since end office switching and use of end office to tandem trunks is not required. Cellular-terminated traffic does include traffic from IXCs passed on by LECs since this traffic is just as costly for

As noted above, other evidence collected on traffic patterns suggests that the amount of interconnected traffic CMRS providers and LECs receive for termination in their busy hour may be less unbalanced than the flow of total traffic. Direct information on the balance of traffic during busy hours was not available, however, and indirect calculations based on limited traffic profile information that is available cannot be made with confidence. We have, however, prepared some calculations intended only to illustrate the magnitude of the adjustment to the total traffic balance that might be supported. These illustrative calculations derive the relative amounts of traffic each carrier receives for termination during its busy hour from total traffic flows under three different sets of assumptions. Each calculation begins with the assumption that total traffic terminated by the cellular system is one third as great as total traffic terminated by the LEC. The three adjustments made and the results of the calculations are as follows:

- Adjustment A: Non-coincident system busy hours for the cellular system and LECs, traffic terminated by the cellular system is at a maximum in the cellular system busy hour, but traffic terminated by the LEC in the LEC busy hour is 85 percent of the maximum hourly flow for terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 39 percent as large as traffic

---

the cellular system to terminate as LEC-originated traffic, and the cellular system does not receive switched access revenue from IXC, although the LEC does.

The mean percent of cellular system traffic in various categories, calculated from the data and estimates provided by CTIA members, is as follows:

Cellular-originated, LEC terminated	60.0%
Cellular-terminated, received from LEC (including IXC traffic)	19.5%
Cellular-originated to IXC, via the LEC	5.1%
Cellular-originated direct to IXC	11.4%
Cellular to Cellular	3.7%

The percentages do not add to 100 percent due to a small amount of unallocated traffic that was reported.

terminated in the LEC busy hour (rather than 33 percent as shown by total traffic data).[19]

- Adjustment B: Total daily terminated traffic is 11.6 times the maximum hourly terminated traffic for cellular-originated, LEC-terminated traffic and 8 times the maximum hourly terminated traffic for cellular terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 48 percent as large as traffic terminated by the LEC in its busy hour.[20]
- Adjustment C: Combines adjustments A and B. Traffic terminated by the cellular system is at a maximum in the cellular system busy hour, but traffic terminated by the LEC in its busy hour is 85 percent of the maximum hourly flow for terminated traffic, and total daily terminated traffic is 11.6 times the maximum hourly terminated traffic for LEC-terminated traffic and 8 times the hourly terminated traffic for cellular terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 57 percent as large as traffic terminated by the LEC in its busy hour.[21]

---

[19] Assume that total traffic terminated by the cellular system is 100 and total traffic terminated by the LEC is 300, and the ratio of total terminated traffic to maximum hourly terminated traffic equals 10 for traffic in both directions. The maximum traffic received in any hour for termination is 10 for the cellular system and 30 for the LEC. In adjustment A, traffic received by the LEC in its busy hour is 85% of maximum hourly terminated traffic, or under these assumptions, 25.5 (i.e., 30 x 0.85). Traffic terminated by the cellular carrier in its busy hour is 10, which is 39% of the 25.5 terminated by the LEC in its busy hour. This adjustment corresponds to Case A in the example discussed earlier.

[20] Assume again that total traffic terminated by the cellular system is 100 and total traffic terminated by the LEC is 300. Assume the ratio of total terminated traffic to maximum hourly terminated traffic is 8 for traffic terminated by the cellular system and 11.6 for traffic terminated by the LEC. This implies the maximum hourly traffic terminated by the cellular system would be 25.9 (i.e., 300/11.6), and the maximum hourly traffic terminated by the LEC would be 12.5 (i.e., 100/8). Assuming the each carrier receives the maximum amount of traffic for termination in its busy hour we obtain the result given, since 12.5 is 48% of 25.9. This adjustment corresponds to Case B in the example discussed earlier.

[21] Begin with the figures in the previous footnote: Maximum hourly traffic terminated by the LEC is 25.9 and maximum hourly traffic terminated by the cellular system is 12.5. Making the further adjustment that traffic terminated by the LEC in its busy hour is 85% of the maximum hourly flow of terminating traffic, the LEC terminates 22.0 units of traffic in its busy hour (25.9 x 0.85); 12.5, the traffic terminated by the cellular system in its busy hour, is 57% of 22.0.

These calculations are no more than illustrative (although each is at least suggested by available information). They do, however, indicate that the balance of total traffic exchanged could be quite different from the balance of traffic imposing capacity costs on the terminating carriers. Starting with total LEC-terminated traffic that is 3 times cellular-terminated traffic, the adjustments reduce LEC-terminated traffic to as little as 1.8 times cellular-terminated traffic. Even such adjusted figures for the balance of traffic tell only part of the story of the balance of costs imposed by interconnection. Those costs depend on the level of capacity cost per minute in each carrier's busy hour as well as the balance of traffic that imposes capacity costs. Before turning to this issue, however, it is important to remember that all the traffic data discussed above are for cellular systems, and reflect current technology and features of cellular system, the current pricing of cellular systems, and the current level of interconnection payments made and received (or not received) by cellular systems.

The next generation of CMRS systems will likely include advances in technology, service features, and pricing options designed to increase traffic per subscriber. Low-power digital handsets, extended battery life, and the capability of receiving and displaying caller number identification will encourage subscribers to use portable terminals throughout the day. Integration of a mobile telephone number with voice messaging will enable subscribers to return calls in instances when they cannot be reached directly. Pricing innovations, such as the free first minute for received calls promoted by the first operating PCS system, can both stimulate total traffic and increase the fraction of minutes originated on the CMRS system. Overall, as CMRS handsets become increasingly good substitutes for fixed telephones, the future traffic patterns of CMRS systems are likely to more closely resemble those of wireline local telephone systems, with the result that the total flow of traffic terminated by LECs and by CMRS systems will come to be more nearly balanced.

An early report lends some support to the proposition that the flow of traffic exchanged will become more balanced between CMRS providers and LECs. The first operating PCS provider, which offers some of the features discussed above, reports that the balance of calls exchanged with the LEC in its area has been about 50-50.[22]

4. *Cost of Added Capacity*

The amount of capacity the LEC and CMRS provider must add to terminate interconnected traffic will be closely related to the amount of interconnected traffic each must terminate during its system busy hour. The costs each must incur, however, will depend on the cost of adding an additional unit of capacity to each network. LEC and CMRS networks obviously differ, and so may the costs of an additional unit of capacity (for example, the capacity to handle an additional 100 minutes of busy hour traffic). LECs must re-size end office switching capacity and tandem switches as well as interswitch trunks where used. CMRS providers also must re-size switching capacity at the MTSO, subdivide cells, and increase backhaul capacity.

IV. Evaluating Compensation Arrangements

Compensation arrangements provide the means by which firms raise revenue to cover costs, but they do more than that. Compensation arrangements affect the overall economic efficiency of the supply of telecommunications services, and the benefits that consumers realize from those arrangements, in at least three ways. First, compensation arrangements affect the level and structure of prices, which in turn act as market signals that shape behavior. Second, compensation arrangements imply mechanisms for monitoring, billing, and collecting for services provided, and those mechanisms may be more or less costly. Finally, compensation arrangements and the pricing of service can affect the development of competition over time. A choice among compensation arrangements should consider each of these three ways in which the compensation arrangement can affect welfare.

The remaining sections of this paper evaluate usage sensitive pricing and bill and keep arrangements from these three perspectives. The next section discusses the static

[22] Joint Comments of the Sprint Telecommunications Venture and American Personal Communications in CC Docket 95-185, March 4, 1996.

welfare properties of these compensation arrangements. The following section discusses the effects of these arrangements on transactions costs. Section VII discusses how compensation arrangements may affect the development of competition and dynamic efficiency.

V. The Efficiency of Price Signals

Prices shape purchasing behavior. Lower prices encourage purchases and higher prices discourage purchases. The level of demand for various products or services in turn directs the allocation of resources and determine how much of which products and services are produced. The policy concern is that the structure and level of prices be set so that they can perform this allocative function efficiently. Prices perform their allocative function most efficiently when their structure and level of prices for a service accurately signal to purchasers the costs of producing that service. It is this function of prices that leads to the prescription, in standard textbook models, that for maximum efficiency price should equal marginal cost.

In this section we discuss how good a job the prices implied by usage sensitive and bill and keep arrangements are likely to do in providing signals that will induce efficient choices by consumers. It may seem obvious that usage sensitive pricing will perform better in this comparison. The simple case against bill and keep is easily stated: Bill and keep arrangements set a price of zero on additional traffic delivered to another network for termination,[23] while most costs of terminating traffic are usage sensitive. Therefore, the simple case concludes, a price of zero sends an inefficient signal since consumers will make additional calls without taking into account the cost imposed by additional traffic. Instead, the simple case suggests that usage sensitive costs should be recovered with usage sensitive prices; price then reflects the cost of additional usage, and will send efficient signals to consumers and the marketplace.

---

[23] As seen above, however, this does not mean that interconnection services taken as a whole are free under bill and keep. Under bill and keep, CMRS providers and LECs each must incur a cost in exchange for receiving interconnection services.

But this argument is too simple. First, it ignores the effects of compensation arrangements on total costs and on dynamic efficiency. Second, a full analysis of the static efficiency of pricing signals is both more complicated and less clearly favorable to usage sensitive pricing than is admitted by this argument. A full analysis should consider both the actual structure of costs as well as the structure of pricing that will be achievable in practice. The efficiency of pricing signals depends on having the structure of prices match the structure of costs, not merely having the average level of prices matching the average level of costs. To begin this analysis, the next section looks at the structure of interconnection costs.

A. The Structure of Interconnection Costs

Interconnection and the exchange of traffic involves at least two kinds of facilities and costs that should be distinguished. Each has its own structure that should be considered in designing prices to recover that category of cost:

1. Costs of facilities dedicated to interconnected traffic. The leading example is the cost of trunks connecting the networks.
2. Costs of the network facilities that each provider uses both to terminate interconnected traffic and to carry and terminate other traffic.

We discuss briefly the structure of the first of these types of cost, and the appropriate structure of prices to recover these costs. We then look in more detail at the cost structure for shared network facilities; these are the interconnection costs most often thought of as usage sensitive.[24]

1. Costs of Dedicated Facilities

The cost of the dedicated circuits connecting CMRS and LEC networks depends on the number and characteristics of the circuits installed, and only indirectly on the amount of traffic carried over those circuits. Costs are driven by the amount of circuit

[24] A third category of possible costs is one-time costs of adapting CMRS or LEC networks to handle or monitor interconnected traffic. Clearly there will be inefficiencies in recovering one-time costs with continuing charges on usage.

capacity in place. Changes in traffic may change the capacity needed, but traffic may also change without affecting these capacity costs if the change in traffic can be accommodated by the capacity already in place. Because these costs do not vary directly with traffic, it will not be efficient to recover them with a simple charge on all units of traffic. As with charges for private lines, and for the same reason, charges to recover these costs should be structured to depend on circuit capacity, not the volume of traffic carried.

The rule for efficient pricing is simple if separate circuits are dedicated full time to carry LEC to CMRS traffic and CMRS to LEC traffic. In this case, the LEC and CMRS provider should each be responsible for the cost of the trunk capacity carrying the traffic it originates. We understand, however, that traffic in both directions often shares the same circuit capacity. The volume of traffic in each direction might then be used to share the cost of this shared capacity, but it will not be efficient to accomplish this with a simple usage charge. First, it will be more efficient to base the sharing of a cost that depends on circuit capacity on relative usage, than to set a per unit usage charge that causes the total amount paid to fluctuate with total usage rather than circuit capacity. Second, it will be more efficient for the sharing of costs to depend on the circuit busy hour usage than on total usage, since it is busy hour usage that will drive the capacity needed.

Finally, it may be efficient to use sharing rules rather than traffic measurements to determine the division of capacity costs. One such rule, often used for trunks interconnecting adjacent LECs, is for each carrier to bear the full cost of the trunks up to some defined "meet point" midway between the networks. Such a rule has the virtue of causing the cost borne by each carrier to vary with the amount of installed circuit capacity, while still potentially saving costs of monitoring usage over the trunks and billing for those costs.

2. *Shared Terminating Network Costs*

An interconnected CMRS or LEC network terminates traffic originated by subscribers to the other network and directed to its own subscribers. Terminating traffic from the interconnected network is mingled with other traffic carried on the terminating network, sharing use of the same switch and trunk facilities, and (in the case of CMRS networks) of cellsite and associated equipment used to establish and maintain radio connections with subscribers. Terminating the traffic imposes a cost on the carrier because an increase in the amount of terminating traffic, like an increase in other traffic carried by the same facilities, can increase the needed capacity.

The costs imposed by terminating traffic are fundamentally costs of increasing capacity, just as the costs of the interconnecting trunk facilities are costs of providing the necessary capacity. The difference is that the capacity of an interconnecting trunk is dedicated to interconnection service, so the cost of that trunk can be identified as a cost of interconnection. Where interconnected terminating traffic shares use of network facilities with other traffic, no identifiable facilities are dedicated to interconnected traffic in general, or to terminating traffic in particular.

Still, the fact that these are costs of capacity determines the structure of shared network costs. Only additional traffic that presses on the capacity of network facilities imposes a cost. Since facilities are sized to provide a specified grade of service during the busy hour, only increases in traffic during the busy hour require investments to increase capacity. It is accurate to say that the costs of the shared network facilities are usage sensitive, but only in the sense that they vary with *some* usage, namely usage during the busy hour. These costs are not sensitive to, or increased by, all increases in traffic. Additional traffic outside the busy hour of a facility, which can be accommodated without increasing capacity, imposes almost no additional costs.

Two further complications in the structure of these costs are relevant. First, it is a simplification to talk only of the system busy hour for the entire network. Different facilities or components of the network can have different busy hours. For example, many portions of local exchange networks carry the most traffic and have their busy hour during the middle of the day. However, the busy hour is in the early evening for some end office switches in residential areas and for the common transport trunks to some residential end offices. The second complication is that the costs of adding capacity to a particular type of facility may vary with the geographic location of the facility, or perhaps the type of equipment at particular locations. These two complications mean that the cost imposed by a minute of terminating traffic does not depend only on whether it occurs in "the" busy hour. The routing of a call will determine how many of the facilities used to terminate that call have their busy hour at that time, and the costs of adding capacity to those particular facilities.

We now turn to the implications of this cost structure for efficient pricing.

B. Matching the Structure of Prices and Costs

The review above shows the basic flaw in the simple argument in favor of usage sensitive pricing. Shared network costs may be sensitive to particular traffic flows, but it does not follow that a uniform price on usage accurately sends a signal of underlying costs. Not all minutes of usage will impose the same costs. This section analyzes in more detail the static efficiency of pricing signals from usage sensitive prices and from bill and keep arrangements. Prices can be usage sensitive, of course, without being based on cost. Any claim that usage sensitive prices send efficient signals of costs, however, will depend on their being based on costs. Therefore the discussion below only analyzes usage sensitive prices based on cost. The first step in this analysis, then, is to see how prices would be derived from cost.

1. Derivation of Cost-based Prices

The following is a very simplified description of how prices for a particular service would be derived from the costs of the set of facilities and related expenses that would provide the capacity necessary to provide that service. The discussion focuses on only a few key steps in the process that are used in the discussion below, and abstracts from many important issues that must faced in deriving unit costs and prices.[25]

---

[25] Among the issues not considered are whether the costs being measured (and on which prices are to be based) are long run or short run costs, and are marginal or service incremental costs. Another issue not considered is the appropriate way of determining the amount of traffic from which the cost will be recovered when capacity is lumpy and more capacity is installed than is immediately needed.

First, determine the costs of the facilities that are used. In the case, the facilities would include various trunks and switches. The investment to create the capacity provided by these facilities include both equipment and installation costs.

Second, the cost of the capacity must be converted to a cost per unit of time. These facilities are long-lived, and their costs will be recovered over their life. Using a depreciation rate and a discount rate, the investment cost is converted to an equivalent cost per unit of time, for example, an annualized cost.

Third, expenses directly associated with operating this capacity are added to arrive at a total cost. For example, annual maintenance costs would be added to the annualized investment cost of the facilities. These steps result in a cost per unit of time.

Fourth, cost per unit time is converted to price by dividing by the number of units of billed usage of this capacity during the unit of time for which costs were calculated.

To give an example, assume that steps 1 through 3 have yielded an total annualized cost of $1 million for the CMRS capacity used both to terminate interconnected traffic originated by LEC subscribers and to carry all other traffic of the CMRS provider. The objective is calculate a price, based on this cost, that will be charged, on completed calls, for each minute of originating usage and each minute of terminating usage of the CMRS network. Say that in a year there are 100 million minutes of originating plus terminating usage; this includes all traffic using these network facilities, not just the termination of interconnected traffic. Dividing $1 million by 100 million minutes of usage yields a price per minute of originating or terminating usage of 1¢ per minute.

If, instead, only usage during a peak pricing period were to be billed, price would be calculated using total usage during the peak period. Say that annual usage during a peak period of 8 AM to 8 PM totaled 50 million minutes. The price per peak period minute would be $1 million divided by 50 million minutes, or 2¢/minute.

2. "Optimal" Pricing

Given that most costs are costs of capacity, what prices would send "optimal," efficient pricing signals? "Optimal" is put in quotation marks because this discussion considers only the effects of pricing signals on static efficiency. Other effects on efficiency, such as the impact on costs of monitoring and billing for usage and on dynamic efficiency and competition, are ignored at this point. We refer to "optimal" pricing throughout this section for convenience, although this pricing is optimized for only one of several relevant criteria.

What prices are optimal in the static sense of sending efficient signals is influenced by both the structure of capacity costs and by the distribution of the demand for calling through the day.[26] Busy hour traffic determines the sizing of network facilities. A first cut at matching price to cost would be to set a price only for usage during the busy hour, while charging a price of zero for all other usage. A price charged only for busy hour usage would be relatively high since it would be paid on only a fraction of total usage. In the hypothetical example above, capacity costs were $1 million, and there were 100 million minutes of total traffic. If we assume the ratio for total minutes to busy hour minutes is 10 to 1, total busy hour traffic is 10 million minutes. A price on only busy hour traffic would be 10¢ a minute, ten times higher than the price of 1¢ that would be charged on all usage, since busy hour traffic is only 1/10 of total traffic.

A price applied only to busy hour usage still may not be theoretically optimal. The relatively high price will depress usage during the single hour it is charged, which may result in some other hour becoming the busy hour. This phenomenon is referred to as "peak shifting." Figure 2 illustrates the point. Panel A graphs the (hypothetical) distribution of traffic throughout a business day; the prices listed for each hour across the top of the graph show that this is the distribution of calling when the price for usage is zero at all times. The busy hour of this distribution is at 2 PM, but traffic is about 90 percent as high during several other hours. Panel B shows the effect of setting a price only for usage during 2-3 PM. Usage declines in that hour while increasing somewhat at

---

[26] For discussions of optimal pricing (in the sense used here), see R. E. Park and Bridger M. Mitchell "Optimal Peak-load Pricing for Local Telephone Calls", RAND, R-3401-1, March 1987, and Bridger M. Mitchell and Ingo Vogelsang, *Telecommunications Pricing: Theory and Practice*, Cambridge University Press, Cambridge, 1991, and the references cited in these works.

other times. The new peak is at 1-2 PM, but traffic is very nearly as high at 10-11 AM, which is a secondary peak of the distribution. Setting price this high for a single hour is not optimal, both because there is no charge for what becomes the busy hour when the peak has shifted, and because traffic has been depressed below capacity during the original busy hour. This means that the high price deters some calling that would not impose a cost.

Further pricing adjustments are required for optimality, and the direction of the needed changes is clear. Price should be somewhat lower during the original busy hour of 2-3 PM, and a non-zero price should be set for traffic during what would become the new busy hour. Charging for usage only during the original peak, and the new peak of 1-2 PM however, could further shift traffic, and create yet another peak.

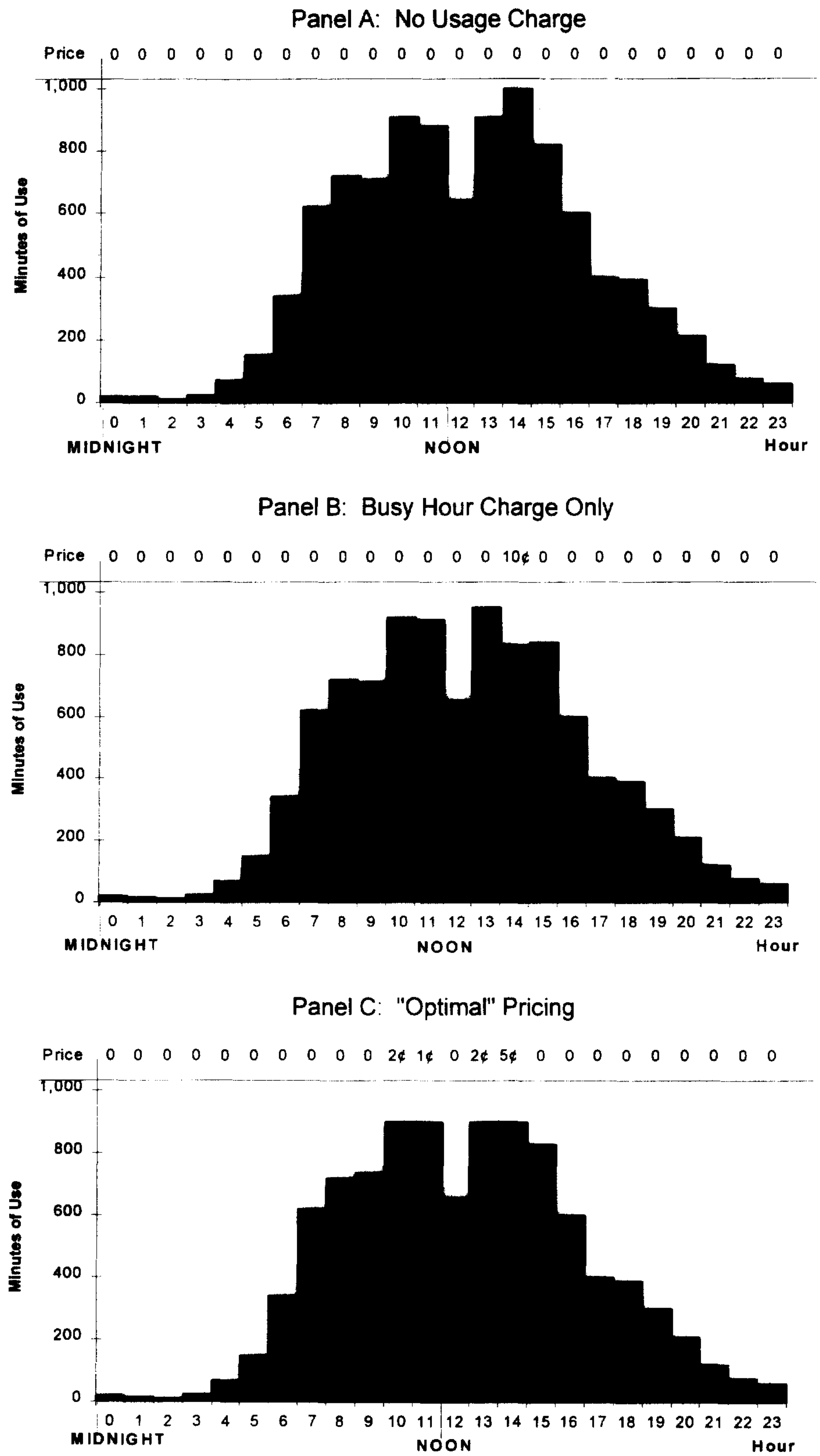
The theoretical solution is to set non-zero prices for several hours, but to set a *different* price for each of these hours.[27] The price to set for each hour depends on both the underlying demand for usage at various times (here manifested by the call distribution when price is zero), and how increased prices at one time will cause usage to shift to other times. Panel C show what such a set of optimal prices might look like, and the resulting distribution of calling. In panel C, non-zero prices are charged for four hours: 10 AM - noon, and 1-3 PM, with prices ranging from 1¢/minute from 11 AM to noon to 5¢/minute from 2-3 PM. Notice that the result of these prices is to make usage the same during each of these four hours; optimal pricing smooths peak usage to create a group of busiest hours in place of a single busy hour.[28] For the other 20 hours of the day outside these busiest hours, price is set at zero.

---

[27] Marcel Boiteux, "La Tarification des Démandes en Pointe", *Revue Générale de l'Electricité* 58: 321-40, 1949.

[28] The other characteristic of the set of optimal prices is that the sum of the prices should equal the marginal cost of a unit of capacity.

Figure 2
Hypothetical Traffic Profiles and Pricing



We can use this "optimal" price structure as a benchmark for comparing the efficiency of pricing signals send by usage sensitive pricing and bill and keep arrangements. Before turning to this, one final point is important. "Optimal" prices have been derived by considering the effects of prices on consumer demand -- that is, on the volume of traffic. Interconnection arrangements, however, set the *wholesale* price paid by the other carrier, not a *retail* price paid by consumers. An additional linkage is needed to apply these results to pricing wholesale service: retail pricing must reflect the structure and level of the wholesale price structure. There are market forces that push to create precisely this linkage. Competition pushes firms to set retail prices based on the level and structure of their costs, including the structure and level of wholesale prices they pay for various inputs. At the same time, retail prices may only approximate the structure of underlying costs, even for competitive firms. Retail prices that more accurately match costs may not occur either because trying to set and collect such prices would increase costs, or because consumers are confused by or otherwise dislike dealing with such complicated pricing. The relationship between wholesale and retail pricing and its significance are discussed in more detail later.

3. Uniform Price per Minute Compared to Bill and Keep

The point of departure for this comparison is that neither a uniform price per minute, nor bill and keep arrangements send pricing signals that are "optimal." This is a comparison of two "suboptimal" pricing structures.

Uniform Price Per Minute

A uniform price per minute never sends quite the right price signal, except by chance. All additional traffic is charged a price, even when network facilities have excess capacity, whereas the correct price signal at such times is zero since additional traffic imposes essentially no additional network costs. Uniform prices also send inefficient signals at all or most times when additional usage does impose capacity costs, because in